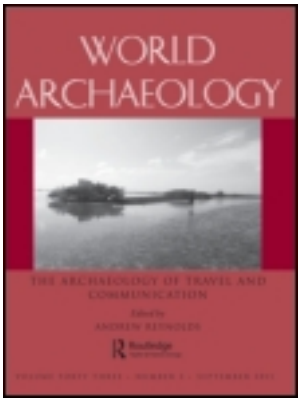

World Archaeology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rwar20

A new Digital Dark Age? Collaborative web tools, social media and long-term preservation

Stuart Jeffrey

Version of record first published: 05 Dec 2012.

To cite this article: Stuart Jeffrey (2012): A new Digital Dark Age? Collaborative web tools, social media and long-term preservation, World Archaeology, 44:4, 553-570

To link to this article: http://dx.doi.org/10.1080/00438243.2012.737579

A new Digital Dark Age? Collaborative web tools, social media and long-term preservation

Stuart Jeffrey

Abstract

This paper examines the impact of exciting new approaches to open data sharing, collaborative web tools and social media on the sustainability of archaeological data. The archiving, reuse and re-analysis of data is often considered intrinsic to archaeological practice, not least because of the destructive nature of excavation. The idea that the pace of adoption of new digital technologies can outstrip the development of the infrastructure required for sustainable access to its outputs, ultimately leading to the loss of data, is sometimes referred to as the 'Digital Dark Age' problem. While strenuous efforts have been made to address this issue, the recent rapid uptake of a new wave of tools to enhance access, promote wider dialogue and gather data has the potential to recreate this problem. This is particularly true because of the volatile technical, legal and commercial contexts in which much of this work takes place. This paper explores these problems, discusses potential changes in the nature of archaeological dialogue and information sharing, and posits solutions that might mitigate a second 'Digital Dark Age'.

Keywords

Social media; archiving; digital preservation; World Wide Web.

Introduction

Archaeology is a discipline whose practice is often predicated on the idea that the information being generated by its practitioners will be available in the long term for reuse and reanalysis. While this can be said to be true of most academic endeavours, it is a particularly potent notion in archaeology due to the intrinsically destructive nature of the excavation process. Merriman and Swain (1999) have extensively discussed the issues surrounding the nature and utility of archaeological archives and a more radical

Routledge
Taylor & Francis Group

World Archaeology Vol. 44(4): 553–570 *Open Archaeology*
 ISSN 0043-8243 print/1470-1375 online
http://dx.doi.org/10.1080/00438243.2012.737579

perspective is offered by Holtorf (2001), questioning the very notion that archaeological evidence is a non-renewable resource. However, the destructive nature of archaeological excavation ensures that considerable emphasis continues to be placed on the preservation and dissemination of the resulting material. In the academic sphere traditional modes of publication such as monographs, books and journal articles still dominate as the favoured means of dissemination, whereas all too often in the commercial sector an unpublished (grey literature) report is all that results from the majority of archaeological interventions. However, both the academic and commercial archaeological sectors generate substantial volumes of data that are not, and cannot be, disseminated via these routes. As this archival material is the source material for the published interpretations, and is the only means of future reinterpretation, it is broadly considered good practice to see that it is archived for reuse in the long term.

Traditionally, documentary and artefactual archival material from archaeological projects has been stored in an assortment of physical locations, typically museums, although there is no universal coverage for this service. Since the advent of *Planning Policy and Guidance 16: Archaeology and Planning* (HMSO 1990) in England and Wales (and similar legislation in Scotland) the volume of material being generated has placed a huge strain on these physical archives and there is a live debate over both selection policy and de-accessioning (see the National Museums Directors Conference report *Too Much Stuff* (Jones et al. 2003)). With the broad adoption of digital technologies, many such archives have also struggled to manage the material being generated in complex digital formats. Debates surrounding the function of the archive, what should be archived and the interpretative nature of the selection process, are equally significant in the digital realm, but there has also been a real need to find efficient technical ways of archiving digital material. As the problems surrounding the long-term preservation of digital archaeological data began to become apparent at the end of the twentieth century the idea of a 'Digital Dark Age' (the title of a talk by Terry Kuny at the 63rd IFLA conference, 1997) was mooted. This broadly refers to the age between the adoption of digital technologies, in our case the production of archaeological data in 'born-digital' formats, and the development of techniques and infrastructure that would allow that material to be kept safe in the long term. The onset of the Digital Dark Age meant that substantial volumes of archaeological data created in digital formats were subsequently lost due to an inability to preserve the material or a lack of foresight in planning for its preservation.

Over time digital archiving procedures have developed and a patchwork of trusted digital repositories emerged. These include the Archaeology Data Service (ADS 2012) in the UK, Data Archiving and Networked Services (DANS 2012) in the Netherlands and Digital Antiquity's 'Digital Archaeological Record' (tDAR 2012) in the USA, although there is by no means global coverage or adequate capacity. Many archaeologists now have access to such digital repositories and for them the Dark Age can be said to have ended as the barriers to digitally archiving archaeological outputs have moved from the realm of the technical to the realm of the financial, political and practical. Virtually all digital outputs could now be kept safe for the long term should that be deemed an appropriate outcome, and if the will and funding were available.

More recently, the model of digital information structures replicating traditional analogue approaches has begun to be challenged as the true potential of the digital

world has begun to be realized. This new approach to information gathering, management and sharing bears little relation to traditional models as practised since the early days of the discipline. It does however offer the potential to benefit everyone involved in the study of the past, from private individuals and community groups to professional field archaeologists and academics. At the forefront of these approaches are social media and collaborative websites that can break the boundaries between professional and non-professional and open the floodgates to the masses of relevant, but untapped, information in the form of user-generated content. While the undoubted benefits of embracing these approaches to archaeological working on the web are well documented (for example, see Cann et al. 2011; Whitcher-Kansa 2011), the potential for a second Digital Dark Age also appears. Once again, the adoption of new techniques and approaches is running ahead of plans or policy to preserve the material generated. This time the problems arising are not so much technical as they are to do with the intersection of online tools with commercial enterprise, the expectation of users with regard to their content and the cost in terms of privacy that participation may imply. A lot of archaeological debate, as well as content creation and sharing can now take place in environments that are open, dynamic and fluid. It may be that these debates take the form of a shared 'flow of consciousness' and there is no expectation or desire that they will ever be migrated to a permanent record. If no such expectation exists then there is no real problem. However, it is not clear that the advocates and users of social, media, content-sharing sites and other online tools have considered fully either the value of what they are engaged in for future use or strategies for ensuring that their content remains available in the future. It is important that, if a *laissez faire* approach is adopted, it is adopted consciously such that what remains available for the long term, if anything, is the result of design rather than accident.

In the sections below, I will discuss why digital data are so fragile, how they are preserved and what the changes in the way archaeologists work on the web might mean for future access and reuse of digital information. I hope to show that the often personal and subjective process of making decisions on what might be appropriate to archive for the future are made particularly pertinent as the commercial infrastructure underpinning social media themselves retains vast amounts of information on their users, the user's content and the user's behaviour.

The first Digital Dark Age

Many of us work with the assumption that some elements of the data we produce in the course of our work in archaeology will be available for researchers to reuse, revise and share in the future. This is especially true as the majority of information we generate is 'born digital', that is, it is created working at a computer or other digital device in the first instance. This assumption is based on the idea that digital data, which are easy to 'back up', copy and share are intrinsically safer than having, say, a single hand-written or typed version of the work. However, the past fifty years or so of digital data management have highlighted a number of fundamental factors that mean that, for long-term access, digital data actually need constant management. This idea might be somewhat counter-intuitive,

but there are numerous factors that mean digital data are easily lost, these include, but are not limited to:

- Data corruption – digital storage media, such as DVDs and magnetic tape degrade over time.
- Media obsolescence – hundreds of previously popular storage formats are now unreadable for practical purposes because the format is obsolete e.g. 5.25-inch floppy discs, Laser discs, Jaz Drives.
- The software used to create or access the data uses proprietary formats or becomes obsolete. Software and file formats change very frequently as technology changes; there is no guarantee that a document created in a word processing application, such as WordStar, will be readable in newer software. This also applies to proprietary (commercially confidential) formats, that is, formats that are created by a specific software package, but are not readable by any other without conversion.
- Inadequate metadata – this problem should, in theory at least, affect only datasets that have been somehow abandoned before they could be prepared for archive. A data file may contain valuable and important information, but without the metadata, that is, the data explaining what it is and how to read and understand it, it may in fact be entirely unusable.

The potential that significant volumes of important work have been generated using software packages or stored on physical media that will become unreadable and inaccessible is a serious one. Nor is a hypothetical threat; there have already been numerous cautionary tales where archaeologists have made assumptions about the security of their data only to discover that they have been lost or will be very expensive to recover (for a good example, see Dunning 2001). Working on solutions to such archiving issues is what the UK's ADS has been engaged in for over fifteen years now.

Preserving digital data

There remains some confusion beyond technical circles regarding the concept of digital archiving. This confusion is not helped by the loose use of the word 'archive' itself in numerous IT-related contexts. Most commonly, the term 'archive' is used to mean either 'back up', i.e. keep a copy of data on a separate mass storage device/service or reformat the data in such a way that they are reformatted/compressed and require de-compression for access (for example a 'tape archive' format).

For the purposes of this paper it is important to be clear that the kind of archiving that is being considered, long-term preservation, implies that the content in question will be usable and accessible on a decadal scale. The question is not whether a particular website will still be around in two or five years' time, the question is whether the data contained therein will be available in twenty-five or fifty years' time. In the fast-moving world of digital technology it is the role of the archivist to think beyond the life of the archaeological project and also beyond the life of a particular web technology and to think about fragility of the work being created over very long periods of time. This may lead to

some difficult questions being asked of early adopters of new technology, particularly in the area of data creation, such as 'your new piece of recording equipment/storage solution/social network seems to have numerous advantages over old ways of working, but, before you commit all your data to it, are you entirely sure it will be recoverable in five years time?' Superficially this question appears to act as a barrier to the adoption of new equipment, techniques or ways of communicating, not least because for many technology users, as opposed to computer scientists, there remains some faith in the permanence, mutability and flexibility of digital data that is not borne out by experience. When software engineers or hardware technologists conceive a new tool, platform or technique the longevity of the digital outputs is rarely foremost in their minds.

One of the most widely acknowledged approaches to the practical matter of preserving digital data for the long term is the Open Archival Information System (OAIS) reference model (CCSDS 2002). OAIS comprises hundreds of pages of guidance and good practice and makes clear the importance of open file formats, data migration, robust and distributed hardware infrastructure and the necessity of discovery, access and delivery systems. It does not, however, detail the specifics of day-to-day practice. Actual digital preservation based on OAIS can be enormously complex. In archaeology, preservation processes may have to deal with literally hundreds of file types, from hundreds of types of devices, hundreds of software packages and the whole gamut of archaeological techniques. In addition, for a digital archive to be credible, for it to attain 'trusted digital repository' status, it must be able to demonstrate fully documented preservation policy and processes as well as having a fully developed long-term sustainability plan. As yet few formal mechanisms exist that act as a form of validation for digital repositories although one such is the Data Seal of Approval (DSA 2012), an international peer review scheme. The ADS was the second UK archive to receive the DSA after the UK Data Archive (for a description of the DSA process, see Mitcham and Hardman 2011).

Social media

While it would be complacent to say that the issue of digital preservation in archaeology is anywhere near resolved, a number of infrastructure solutions have begun to appear and are already having a significant impact on archaeological practice in their respective countries. However, we are now entering a very different world from one where digital technologies are simply being used to replicate the work flows and practices of the analogue age, and new archiving issues arise as a result. The true potential of digital technology has been revealed in the profoundly interconnected world of the World Wide Web. In this arena data are not simply presented for passive consumption in the forms and structures that they have been for hundreds of years, but paths to discovery, sharing, reuse, enhancement and recombination have been hugely accelerated and simplified. This has occurred to such an extent that ways of working undreamt of at the turn of the millennium are now commonplace. The impact of social media in particular has begun to change the traditional network of relationships between archaeologists, their study material, their outputs and their perceived audiences. These changing relationships both

reflect and effect broader social changes as digital technologies have evolved from being facilitators of social practice to become the dominant engine of change in that practice.

The provision of research infrastructure (including archives) has traditionally been the responsibility of academic institutions and government departments. Facilities such as libraries, laboratories, meeting rooms and conference facilities have all helped support the work of the researcher and the research community and access to these facilities has been mediated by the political and social structures from which these agencies spring. A profound shift has taken place recently with regard to research infrastructure for data sharing, collaboration and dissemination. The most useable, most used and most technically sophisticated environments for communicating with, working with and sharing information with like-minded people are now often situated in the commercial social media sector of the World Wide Web.

Given the vast number of potential social media sites and channels available for use it is difficult to analyse them critically as if they were a single coherent set of tools. This is especially true as each site or service has unique terms and conditions, rules for participation and access as well as policies regarding appropriate content, ownership of content and storage of content. Any attempt at a definitive list of the type of tools would be a huge undertaking, as well as being out of date very quickly indeed. Nevertheless, a short list of commonly used tools, after Cann et al. (2011), is given below, to illustrate the range of services now on offer:

- Communication (URLs for each site are given in the references):
    - Blogging – Blogger, Live Journal, WordPress
    - Microblogging – Twitter, Yammer
    - Location – Foursquare, Gowalla
    - Networking – Facebook, Linkedin, MySpace, academia.edu
- Collaboration:
    - Conferencing – GoToMeeting, Skype
    - Wikis – PBworks, Wetpaint, Wikia
    - Bibliography – CiteULike, Mendeley
    - Documents – Google Docs, Dropbox, Zoh
- Multimedia (sharing):
    - Images – Flickr, Picasa, SmugMug
    - Video – Viddler, Vimeo, YouTube
    - Presentation sharing – Scribd, SlideShare
    - Virtual Worlds – OpenSim, Second Life.

Given the nature of these tools many of them can be, and frequently are, used together e.g. video may be hosted on a streaming service such as Vimeo, but embedded in a blog created

in WordPress. Similarly, many applications could be said to fall into more than one of the broad categories in the list above.

The adoption of these tools and platforms is likely to accelerate in the coming decades as the hardware choices available impact on the software people choose to use. This is because the advent of mobile computing, via devices such as tablet computers, iPads and smart phones may well represent a revolutionary step in the development of the IT environment. It is perfectly feasible that, just as mobile phones have begun to make land lines near obsolete (or have seen widespread adoption in areas that never had a static telephone infrastructure), mobile computing, with its reliance on software as a service and cloud-based storage, will ultimately entirely supplant our current desktop-computing culture. It is agile and innovative commercial organizations that are responding fastest to this development: GoogleDocs is relatively easy to use on an iPad for word processing, MicroSoft Word is not.

The second Digital Dark Age

The enormous current and potential benefits of social media tools are discussed at length elsewhere (see, for example, Cann et al. 2011; Procter et al. 2010). These benefits obviously include improvements in the way people are able to work with each other, share ideas and information. They can also effectively facilitate communication between academic, professional, community and public archaeology, greatly enhancing a project's potential impact. However, there are clear implications for the way that digital data are managed, particularly the relationship between social media content and the archaeological archive. What is clear is that we cannot unthinkingly assume that the long-term preservation of digital data is a problem that either has been solved or is likely to be solved by new applications or web tools. Where data are being created using locally controlled infrastructure, hardware and software, it is possible to engage the services of a digital archive which can undertake to keep this material safe and to make it freely and openly available into the future. This is the model that has developed, slowly and with some difficulty, over the last twenty years or so. An archaeologist generates data and digital archives such as the ADS, DANS and tDAR accession the material and make it available for the long term. This pattern has begun to change, the infrastructure in which archaeologists communicate, collaborate and create datasets is now often no longer under their direct control. This change in practice has led to two changes in traditional academic behaviour: one is engagement with the unmediated commercial web space, where little user control can be exercised over the environment, and the other is participation in environments where distinctions between professional and personal activity break down. The so-called Digital Dark Age arose when the adoption of new technologies overtook the development of infrastructure and policy enabling the preservation of the new digital outputs. This new wave of technological change could potentially see the same temporary situation arise, a second Digital Dark Age.

The main lesson that should be drawn from previous experience in digital archiving is that enthusiasm for adopting new approaches should be tempered with some consideration of the longevity of the outputs.

For the vast majority of everyday interaction via social media and other online tools longevity is simply not an issue. As discussed above the process of selecting material for archiving is well understood to be an interpretative process and inevitably the subjective nature of interpretation means that guidance on what is appropriate for archiving will always remain guidance (however strongly stated), as no universal or definitive statement on what warrants archiving is possible. For content created or hosted in social media the same applies. It is also true that special consideration needs to be given to the desirability of archiving any material from social media at all – not least because the social media sphere is often considered a 'free', 'open' or even 'neutral' space in which transient, free form engagement takes place and whose participants are actively reacting against traditional, formal, mediated and permanently recorded modes of discourse. In the sections below, I hope to show that, while this is generally true, it is not always so. There are some senses in which the commercial social media sphere does constitute a permanent record, but one which is outwith the user's control and there are also some instances of social media use where long-term preservation of content is in fact entirely appropriate.

If there is any likelihood of content created and/or hosted in social media being considered valuable, by whatever criteria, then the appropriateness of the tools being used should be considered in this light. When engaging with archaeological discourse in social media environment, or online shared working platforms, users have to balance the benefits of using these modes of communication against several important points:

- Users are not directly in control of their content because their 'contract' with the platform is unlikely to guarantee perpetual access. Raising the question: how long do you wait before deciding content is worth keeping, for example, by extraction via Google's 'Data Liberation' tools? This is a particularly important question to ask if short- or medium-term storage capability is being used via sites like Flickr, Scribd and Dropbox. The content is almost certainly safe in the short term, in that these platforms have reliable infrastructure (although they may not be appropriate for sensitive data (Bott 2011)). However they are not an archival solution in the digital repository-specific meaning of the term.
- Accidental data loss does occur in a commercial context, for example, the catastrophic failure of the Ma.gnolia social bookmarking site, which shook many people's faith in web-based storage in general (Calore 2009). Another example of data loss on a smaller scale from Flickr is discussed by Wauters in a Techcrunch article entitled 'Flickr accidentally wipes out account: five years and 4,000 photos down the drain' (2011) .
- Access to a service can be curtailed for a number of reasons: abuse by other users (e.g. a legal injunction over piracy at file-sharing site Megaupload resulted in large numbers of blameless users losing access to their data and even facing deletion), the introduction of charging, the introduction of new, unwelcome functionality, such as intrusive registration processes or information-sharing regimes, e.g. Facebook (Bankston 2009; Sutton 2012). The service is sold or blocked (e.g. Twapperkeeper was both blocked from Twitter and then sold (Kelly 2011; Yin 2011)), it changes its conditions of use (Facebook and Google (Bankston 2009)), it is declared obsolete (e.g. Geocities, where many years' worth of content was put at risk and which was

the centre of an emergency archive project (GeoCities Archive Team 2012)), or the service simply goes out of business (e.g. SixDegrees.com). Commercial might and marketing power are no guarantee that services will be maintained if they are not commercially viable, for example, Google's microblogging site Google Buzz was closed shortly after launch, although some features were picked up by Google + and content is currently still recoverable (Google + 2012).

- Tools allowing the extraction of data from services may not exist or may be very limited in the formats and structure they can handle (this has most significance when large amounts of data need to be extracted from a site, for example large volumes of video clips from a YouTube channel or 3D models from Second Life).
- Each platform for dialogue has its own self-selecting audience. It is in the interest of the platform or service to attract and keep users, so links between platforms are often constructed with this in mind. Should engagement take place over multiple social media channels? This not only has an impact in time for the researcher, but raises the potential for useful information being unhelpfully distributed over multiple social media platforms.

There is a clear premium associated with the word 'open' when attached to software, platforms and data sources. The word itself is a somewhat loaded term of course and there is no settled definition what it actually means. Truly open content or a truly open platform requires nothing in exchange from the user and places no restrictions on how the content or platform is subsequently used. In practice social media platforms, while often 'open' and 'free', almost universally require registration and the information gathered through the registration process is not without value; the registration is, in fact, a transaction even if there is no charge levied. Personal information about users and the behaviour of users within a social media environment is well recognised as having commercial value, even in anonymized form. As Doctorow points out in his article 'The curious case of internet privacy' (2012), these data are in fact one of the main planks in the business model of many ostensibly 'free' sites. Of course direct advertising is a mainstay of free-to-use services and some social media platforms, such as YouTube and Facebook, which built up substantial user bases have adopted advertising as means of 'monetizing' that user base and generating revenue. It is the ability to tailor this advertising to specific users' interests that imparts much of the value to the personal information captured at registration (or by other means, such as via Google's StreetView programme (O'Connor 2012)). Bear in mind that cross-registration, or 'social login', whereby one site uses the registration and identification systems of another (e.g. logging into Scribd using a Facebook login) effectively ties one account to the other, creating a single, trackable, online identity. Another business model is where an organization focuses on generating a large, and importantly, reliant, user base, and then introduces charging for premium services, or even services that were initially free. An example of this was the introduction of charging for volume use of the GoogleMaps API, a previously free service that had come to be integral to the work of a large number of users (see Arthur 2011). While each individual researcher might be entirely comfortable with the arrangement that they enter into with social media, in terms of advertising, information sharing and privacy, working solely in this way may exclude others who are not.

User-generated content and user expectations

So far, I have discussed the activity of the individual researcher and the points they might wish to consider with regard to the archiving of their own material held in social media platforms. However an increasingly popular use of social media takes advantage of its capability, not just for collaborative work with peers and colleagues, but for engaging wider audiences such as community groups or interested members of the public. This can take place via the use of Wikis or public participatory GIS, and can range in scope from a discourse on a particular monument or landscape to a general invitation to contribute stories about the past (e.g. HistoryPin 2012). Such projects offer the opportunity to engage directly with people outside the cultural heritage professions and consider their perceptions of the past, the value they place on cultural heritage and the ways that it has been used to inform the construction of personal, local and national identities. While being of significant use in these (and other) research contexts, a separate set of issues arises in using social media to engage with these audiences, particularly if the engagement is interactive and the wider audience is encouraged to create and submit content. Such content could be a simple text entry, digital images or even sound and video. User expectations of the ultimate fate of this material should be managed carefully. There is an ethical imperative to ensure that contributors are aware of the longevity of the content they produce, for instance, if a decision has been taken that the user-generated content will not be archived and/or will not be made available beyond the life of the particular project. Similarly, if the decision has been made that material will be available for as long as the social media platform being used allows, then this should be made apparent at the point users choose to engage. This does not mean that an archiving strategy cannot be developed after content has been gathered, but users may reasonably have some expectation that their content is being collected with the objective that it is preserved for the long term. It is not too difficult to imagine a situation where the only version of a text, image or video is the one submitted to a project because a private individual assumed it would be safe. For this type of use of social media more than any other it is important to consider the likely archival strategy at the outset. It is not wise to assume that other participants in a collaborative project or interactive content-gathering exercise share your perceptions of the future utility, ownership and permanence of the content being created.

Selection and retention

As with all archiving processes, physical or digital, selection and retention (and disposal) play a vital role in how the archive is constructed. For the museum sector, focusing mainly on non-digital material, Merriman (2008) makes a fascinating case for the benefits of disposal and 'forgetting'. For financial and practical reasons what is deposited in an archive is normally a subset of the original material. Simple examples of this process might be the discarding of blurred or uninformative site photographs and the archiving only of those considered useful. The selection process may be motivated by the aforementioned financial and practical considerations, but it is also undoubtedly an interpretative process. Decisions are made on the value of each potential archive element based on a prediction of

what will be important in the future; some predictions are well founded, but there are numerous examples of datasets that were discarded but were later discovered to have been potentially highly valuable or where the most unlikely datasets turn out to have significant reuse value in way that was entirely unpredictable. A very well-known example of this is nineteenth-century whaling records being reused in modern climatology research (see de la Mare 2009). Despite this, even where finance is not an issue, normally not everything from a project is deemed appropriate or desirable for archive.

Selection and retention are equally important in an online environment. Not every online conversation on the topic of archaeology should, or could, be archived. The space for transient conversations, where opinions can be expressed and then forgotten, new ideas can be discussed without the fear that they will return endlessly to haunt one, should be cherished. Two important points to note arise from this. First, and quite obviously, the vast majority of interactions in social media or the broader web space simply do not warrant archiving or should not be archived for other reasons. Second, a record of activity which has taken place in the web space may well in fact be kept, not necessarily archived in the most correct use of the word, but retained. It may be retained by the hosting application, because the ability to engage in mining of these data, tracking interactions, connections and 'mentions' in a social web space is something that itself has value, for example where the Twitter archive is sold to marketers (Barnett 2012). So a strange and perhaps unexpected situation arises. Decisions about what is retained from the web are not being made by the content creators, but they are being made by the owners of the application managing the content. This has the effect of eroding what has been called 'the right to be forgotten' (Bright 2012; Fiveash 2012), a vital component of a platform for free and open debate, as well an important privacy issue. It also undermines the control that content creators have over their content, although it has been argued by Orlowski (2012) that this situation is a temporary one and that it is technically possible to regain control over data retention and privacy. It should be noted that, although many content creators may not actually desire any control over their content, this is not a universal position. Of course, in contrast with the archiving processes that a trusted digital repository might be engaged in, data retention by commercial organizations is entirely self-serving. The data will be retained for as long as they have a value to the owner of the platform or application, and as long as the host continues to exist. It is a matter of concern for many people that a broad range of data on users and their behaviour are retained as a matter of course by commercial organizations; this issue is particularly potent in countries where state security organizations have easy legal access to such data and an interest in scrutinizing the behaviour of its citizens. Long-term preservation of data for the purposes of future research by the actual content creators, however, is simply not an area of business that social media companies are currently focused on.

Referencing

An interesting element in the debate on the archiving of social media is the difficulty of referencing for academic purposes. If an important fact, link or piece of data appears in the world of social media, how does one refer a colleague to it (or find it again for oneself)?

The problem of permanent unique references is essentially as old as the internet; the use of URLs to reference individual web pages or elements therein has been proven to fail on two fronts regarding longevity. First, URLs need constant maintenance as the systems around them are updated and the files they refer to disappear or are relocated. As many as half of all hyperlinks links on the World Wide Web no longer work after just ten years, resulting in dead links or what has been referred to as 'link rot' (see Wikipedia's 'Link Rot' entry 2012). Second, large amounts of web content are actually generated dynamically, i.e. the page in question is not a static HTML file with a unique address in the first instance. With social media there are often quite easy ways to reference a particular post or piece of information, but to quote Chelsea Lee, Senior Manuscript Editor of the American Psychological Association Journals, on the citation of Twitter and Facebook:

> Because online social media are more about live updates than archiving, we don't know if these status update pages will still be here in a year, or 5, or 20 years. So if you are writing for publication, it may be prudent to self-archive any social media updates you include in your articles.
>
> (http://blog.apastyle.org/apastyle/2009/10/how-to-cite-twitter-and-facebook-part-i.html)

Ironically, as this article's citation to the above quote is a URL to a blog post, the likelihood of that URL working in ten or twenty years is actually quite low. Lee states that it is prudent to self-archive; however, more sensible would be to ensure that the relevant content is archived as part of a dataset lodged with a repository that offers a permanent unique reference.

There are number of schemes that tackle this issue sensibly, most based on the 'handles' system. The ADS, for example, uses a Digital Object Identifier (DOI) system. This allows collections of data or individual data files to be allocated a URL that will not change irrespective of changes to the physical location of the files in question. Organizations that support the 'minting' of DOIs, such as the DataCite (2012) project, have made a commitment to ensure that the DOI URL will always resolve correctly; this is not a trivial commitment and as a result organizations participating in DataCite are carefully vetted. In addition to the advantage of some level of permanency being introduced into the citation, this also has the benefit of not requiring the user to register and participate in the social media platform in question to access the referenced data. Given the potential political sensitivities and privacy issues raised by some registration processes as discussed above (e.g. cross-registration), access to referenced content via a third party archive should largely mitigate these concerns. This assumes that the archive itself does not engage in intrusive, compulsory registration processes, user tracking and, particularly, information sharing with market intelligence companies.

Archiving content developed using social media.

Much work has started in the archiving world looking at the development of systems and policies that facilitate the archiving of social media content. Two ongoing projects

exemplify how the use of social media can be integrated with good archival practice. The ADS has worked with both the Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2012) and the collaborative social media project Day of Archaeology to act as an archive for the content generated via their respective use of web tools. The CAA held in the University of Southampton in 2012 made extensive use of the microblogging site Twitter to communicate information to attendees and to offer a platform for attendees to share comments and further information throughout the conference. The resultant set of 'tweets' contains some valuable data on the attitudes and knowledge base of CAA attendees that could reasonably be assumed to be of potential interest to future researchers either inside or outside the archaeology domain. Rather than rely on their long-term availability via the original web tool, Twitter, the CAA organizing committee, working with the 'Social Media in Supporting Live Events' project (SMiLE 2012) based at the University of Southampton's Digital Humanities group, took the decision to engage actively in archiving this material with the ADS (and see Harris and Beale 2012). Given the text-based nature of 'tweets' the technical aspects of preserving this material and making it searchable are in fact relatively straightforward.

The Day of Archaeology project (2012), first launched in 2011, is a groundbreaking project organized by a collective of interested archaeologists led partly by Lorna Richardson, a PhD candidate at University College London. The project encouraged blog posts, including images and video from archaeologists from all around the world on one particular day. In this way a 'snapshot' of archaeological activity was built up covering all sectors, including academic, commercial, fieldworkers, specialists, students and curators. Using the web blogging tool WordPress as a core application, an exciting body of work was created. It is clear that the snapshot created on that day in July 2011, as well as fulfilling its role of information sharing and community building among members of the profession, could also well become a valuable document for the historians of the archaeological discipline in the future. For the 2012 Day of Archaeology, a repeat of the 2011 event, the ADS has supported the project by offering a long-term archive for the material. The ADS is not replacing the Day of Archaeology web presence; rather they will engage in extracting the content and making it available as a separate entity. Will WordPress and the Day of Archaeology site itself still be available in five or ten years, will the video formats used still be readable by software in 2030? Obviously it is impossible to answer these questions, but the project organizers have mitigated this concern by passing the responsibility for long-term access to content to the ADS. One important point to note, although not covered in this paper, is that careful consideration needs to be given to the issue of privacy, copyright and IPR when archiving such material. In both the projects mentioned above explicit permission was sought for archiving from the original data creators, the tweeters and the bloggers.

For the CAA and the Day of Archaeology a reasonable case can be made regarding the potential future value of the information generated via social media relating to discrete projects. Twitter traffic associated with a conference probably comes under the category of ephemera; however, there are already a number of examples of archival practice being employed for large Twitter datasets relating to major events such as the 2011 political upheavals in Egypt and Libya (Small et al. 2012) as well as discussion on the extreme fragility of event-based data in social media (Salaheldeen and Nelson 2012). While

obviously less historically significant than the upheavals in the Arab world, public response via blogging and microblogging sites to important archaeological discoveries or contentious sites could represent both a useful resource for the analysis of public perceptions and for evidence of research impact. The Day of Archaeology blog posts are a grass-roots expression of what archaeological work is actually like at a particular place and time and numerous potential uses can be imagined, from social history research to studies of archaeological practice. Both projects demonstrate that, with foresight, relatively straightforward processes can be put in place to archive material deemed to be important. In this instance engagement with a trusted digital repository, the ADS, will allow the data creators to archive material successfully with the minimum of fuss. It should be understood that this approach is not always appropriate; in particular, it is hard to argue for some kind of wholesale preservation of any social media content with an archaeological theme, for example the ADS does not archive its own Twitter activity and an active debate exists around the wisdom of the Library of Congress's project to archive all Twitter traffic. However, the selection and retention issues outlined in the section above should always be considered, whether your data are derived from an excavation or from a user-generated content project or from an online debate on Twitter or Facebook.

Conclusion

It is a testament to the forward thinking and innovative nature of much archaeological research that social media have begun to be adopted with such enthusiasm and the observations on the complexities they pose for long-tem preservation should be considered in this context. It has not been possible in this overview to focus on the specific issues posed by each platform, service or tool. Furthermore, technical aspects of archiving highly interactive datasets have not been discussed at all (for some discussion of these issues, see Jeffrey 2010). It should also be clear that the first 'Digital Dark Age' was in fact temporary. Good data management and archival practices are catching up with the digital technologies being employed. Similarly, a second 'Digital Dark Age', associated with new technologies, should also be temporary. However, because the issues influencing the way that social media can and should be archived are not trivial, efforts should be made to avoid a mismatch between data creation and management practice and long-term preservation. As has already been stated, for the overwhelming majority of interactions between archaeologists and the world of social media the concerns of long-term preservation simply do not arise. It is only in quite specific uses of these tools that this needs to be a consideration at all. These cases are likely to include where social media are being used as a short-term storage and/or dissemination solution for images, texts or records relating to an archaeological work, where discourse of potential future value is taking place and where community contributions (user-generated content) are being gathered. The issues discussed in earlier sections of this paper (ranging from control, permanence and terms of access to user expectations, IPR, selection and retention and referencing) apply to these uses of social media, just as they apply to data created via more traditional digital means.

How then to avoid a second digital Dark Age? Where we migrate our communication, collaboration and content creation and management to new services, we should acknowledge that, in terms of long-term sustainability, the work we do here is as fragile, arguably more fragile, than the work done on traditional IT infrastructures. Individual researchers or joint archaeological projects routinely take responsibility for decisions regarding the deletion or selection and retention of digital outputs created in the traditional fashion and, if there is to be an archive, they decide how this will be managed. Ideally, this should happen at the planning stage of any proposed research programme when provision should be made for the long-term care of the expected outputs. I suggest that archival planning should include, where appropriate, outputs generated in the social media sphere. The migration of activity from, say a university or company infrastructure, conferences and workshops into the world of web tools and community engagement does not change the fact that data are being created and important discourses are taking place. There remains a responsibility to ensure that data which, by whatever criterion, warrant archiving are actually archived. As discussed above, this is not an argument for archiving the everyday coming and goings in the sphere of social media, particularly social networks, where archiving for academic purposes is generally neither appropriate or desirable, and data retention by the service providers themselves could have undesirable consequences. Rather, it is an argument for the decisions on archival policy being arrived at through active consideration of the issues social media raise rather than by default. There is a world of difference between actively deciding that archiving is not important and accidentally discovering that it is not possible.

References

academia.edu. http://academia.edu (accessed August 2012).

Archaeology Data Service (ADS). http://archaeologydataservice.ac.uk (accessed August 2012).

Arthur, C. 2011. Google puts a limit on free Google Maps API: over 25,000 daily and you pay. *The Guardian*, Technology Blog. Available at: http://www.guardian.co.uk/technology/blog/2011/oct/27/google-maps-api-charging (accessed August 2012).

Bankston, K. 2009. Facebook's new privacy changes: the good, the bad, and the ugly, electronic frontier foundation. Eff.org. 9 December. Available at: https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly (accessed August 2012).

Barnett, E. 2012. Twitter sells tweet archive to marketers, *The Daily Telegraph*, 28 February. http://www.telegraph.co.uk/technology/twitter/9110943/Twitter-sells-tweet-archive-to-marketers.html (accessed August 2012).

Blogger, http://www.blogger.com (accessed August 2012).

Bott, E. 2011. Sorry, Dropbox, I still don't trust you, *ZDNet*. Available at: https://www.zdnet.com/blog/bott/sorry-dropbox-i-still-dont-trust-you/4173 (accessed August 2012).

Bright, P. 2012. Europe proposes a 'right to be forgotten'. *ars technica*. Available at: http://arstechnica.com/tech-policy/2012/01/eu-proposes-a-right-to-be-forgotten/ (accessed August 2012).

Calore, M. 2009. Ma.gnolia suffers major data loss, site taken offline. *Wired Magazine*. Available at: http://www.wired.com/business/2009/01/magnolia-suffer/ (accessed August 2012).

Cann, A., Dimitriou, K. and Hooley, T. 2011. Social media: a guide for researchers. Research Information Network. Available at: http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/social-media-guide-researchers (accessed August 2012).

CiteULike. http://www.citeulike.org/ (accessed August 2012).

Computer Applications and Quantitative Methods in Archaeology (CAA). http://caaconference.org/ (accessed August 2012).

Consultative Committee for Space Data Systems (CCSDS). 2002. Reference model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January. Available at: http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed August 2012).

Data Archiving and Networked Services (DANS). http://www.dans.knaw.nl/ (accessed August 2012).

Data Seal of Approval. http://datasealofapproval.org/ (accessed August 2012).

DataCite. http://datacite.org/ (accessed August 2012).

Day of Archaeology. http://www.dayofarchaeology.com/ (accessed August 2012).

de la Mare, W. K. 2009. Changes in Antarctic sea-ice extent from direct historical observations and whaling records. *Climatic Change*, 92: 461–93.

Digital Antiquity. www.digitalantiquity.org/ (accessed August 2012).

The Digital Archaeological Record (tDAR). http://www.tdar.org/ (accessed August 2012).

Doctorow, Cory. 2012. The curious case of internet privacy. *Technology Review*. Massachusetts Institute of Technology. Available at: http://www.technologyreview.com/news/428045/the-curious-case-of-internet-privacy/ (accessed August 2012).

Dropbox. http://www.dropbox.com (accessed August 2012).

Dunning, A. 2001. Excavating data: retrieving the Newham Archive. Available at: http://www.ahds.ac.uk/creating/case-studies/newham/index.htm (accessed August 2012).

Facebook. http://www.facebook.com (accessed August 2012).

Fiveash, K. 2012. Google exec questions Reding's 'Right to be forgotten' pledge in. *The Register*. Available at: http://www.theregister.co.uk/2012/01/26/google_exec_criticises_right_to_be_forgotten_proposal/ (accessed August 2012).

Flickr. http://www.flickr.com (accessed August 2012).

FourSquare. https://foursquare.com/ (accessed October 2012).

GeoCities Archive Team. http://www.archiveteam.org/index.php?title=GeoCities (accessed August 2012).

Google+. https://www.google.com/+ (accessed August 2012).

GoogleDocs. http://docs.google.com (accessed August 2012).

GoToMeeting. http://www.gotomeeting.com (accessed August 2012).

Gowalla. https://en.wikipedia.org/wiki/Gowalla (accessed October 2012).

Harris, L. and Beale, N. 2012. If you don't have social media, you are no one: how social media enriches conferences for some but risks isolating others. SMiLE blog post. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2012/05/23/social-media-enrich-but-isolate/ (accessed August 2012).

Her Majesty's Stationery Office (HMSO). 1990. *Planning Policy and Guidance 16: Archaeology and Planning*. Available at: http://www.communities.gov.uk/documents/planningandbuilding/pdf/156777.pdf (accessed August 2012).

HistoryPin. http://www.historypin.com/ (accessed August 2012).

Holtorf, C. J. 2001. Is the past a non-renewable resource? In *Destruction and Conservation of Cultural Property* (eds R. Layton, P. G. Stone and J. Thomas). London: Routledge, pp. 286–94.

Jeffrey, S. 2010. Resource discovery and curation of complex and interactive digital datasets. In *Revisualizing Visual Culture* (eds H. Gardiner and C. Bailey). Burlington, VT, and Farnham: Ashgate.

Jones, M., Bullick, S., Crump, M., Merriman, N., Swain, H. and William, E. 2003. *Too Much Stuff*. National Museums Directors Conference Report. Available at: http://www.nationalmuseums.org.uk/media/documents/publications/too_much_stuff.pdf (accessed August 2012).

Kelly, B. 2011., Responding to the forthcoming demise of TwapperKeeper. UK Web Focus. https://ukwebfocus.wordpress.com/2011/12/11/theforthcoming-demise-oftwapperkeeper/ (accessed June 2012).

Kuny, T. 1997. A Digital Dark Age? Challenges in the preservation of electronic information. Paper given at 63RD IFLA (International Federation of Library Associations and Institutions) Council and General Conference. Available at: http://archive.ifla.org/IV/ifla63/63kuny1.pdf.

Linkedin. http://www.linkedin.com (accessed August 2012).

Live Journal. http://www.livejournal.com (accessed August 2012).

Mendeley. http://www.mendeley.com (accessed August 2012).

Merriman, N. 2008. Museum collections and sustainability. *Cultural Trends*, 17(1): 3–21. Available at: http://dx.doi.org/10.1080/09548960801920278

Merriman, N. and Swain, H. 1999. Archaeological archives: serving the public interest? *European Journal of Archaeology*, 2(2): 249–67, doi:10.1177/146195719900200206.

Mitcham, J. and Hardman, C. 2011. ADS and the Data Seal of Approval: a case study for the DCC. Available at: http://www.dcc.ac.uk/resources/case-studies/ads-dsa (accessed August 2012).

MySpace. http://www.myspace.com (accessed August 2012).

O'Connor, R. 2012. Google is evil. *Wired Magazine*, June. Available at: http://www.wired.com/business/2012/06/opinion-google-is-evil/ (accessed August 2012).

OpenSim. http://opensimulator.org (accessed August 2012).

Orlowski, A. 2012. Habeas data: how to build an internet that forgets. *The Register*, June. Available at: http://www.theregister.co.uk/2012/06/11/habeas_data_fighting_data_expiry/ (accessed August 2012).

PBworks. http://www.pbworks.com (accessed August 2012).

Picasa. http://picasa.google.com (accessed August 2012).

Procter, R., Williams, R. and Stewart, J. 2010. If you build it, will they come? Research Information Network. Available at: www.rin.ac.uk/system/files/attachments/web_2.0_screen.pdf (accessed August 2012).

Salaheldeen, H. and Nelson, M. 2012. Losing My Revolution: How Many Resources Shared On Social Media Have Been Lost? Theory and Practice of Digital Libraries (TPDL) 2012 arXiv:1209.3026 [cs.DL].

Scribd. http://www.scribd.com (accessed August 2012).

SecondLife. http://secondlife.com (accessed August 2012).

Skype. http://www.skype.com (accessed August 2012).

SlideShare. http://www.slideshare.net (accessed August 2012).

Small, H., Kasianovitz, K., Blanford, R. and Celaya, I. 2012. What your tweets tell us about you: identity, ownership and privacy of Twitter data. *International Journal of Digital Curation*, 7(1). Available at: http://www.ijdc.net/.

SMiLE project. http://caaconference.org/caa2012/events/smileproject/ (accessed August 2012).

SmugMug. http://www.smugmug.com (accessed August 2012).

Sutton, M. 2012. International reactions to Google's new privacy policy. Electronic Frontier Foundation. Eff.org. Available at: https://www.eff.org/deeplinks/2012/03/international-reactions-googles-new-privacy-policy (accessed August 2012).

Twitter. https://twitter.com (accessed August 2012).

Viddler. http://www.viddler.com (accessed August 2012).

Vimeo. http://vimeo.com (accessed August 2012).

Wauters, R. 2011. Flickr accidentally wipes out account: five years and 4,000 photos down the drain. techcrunch.com. Available at: http://techcrunch.com/2011/02/02/flickr-accidentally-wipes-out-account-five-years-and-4000-photos-down-the-drain/ (accessed August 2012).

Wetpaint. http://wiki.wetpaint.com/ (accessed October 2012).

Whitcher-Kansa, S., Kansa E. C. and Watrall, E. (eds). 2011. *Archaeology 2.0 and Beyond: New Tools for Collaboration and Communication*. Los Angeles, CA: Cotsen Institute of Archaeology. Available at: http://escholarship.org/uc/item/1r6137tb.

Wikia. http://www.wikia.com (accessed August 2012).

Wikipedia: Link Rot. https://en.wikipedia.org/wiki/Link_rot (accessed August 2012).

WordPress. http://wordpress.org (accessed August 2012).

Yammer. https://www.yammer.com (accessed August 2012).

Yin, S. 2011. Twitter cracks down on another third-party app, TwapperKeeper. *PC Magazine*. Available at: http://www.pcmag.com/article2/0,2817,2380784,00.asp (accessed August 2012).

YouTube. http://www.youtube.com (accessed August 2012).

Zoho. http://www.zoho.com (accessed August 2012).

Stuart Jeffrey has been with the Archaeology Data Service, an open access digital archive for archaeological research outputs based at the University of York in the UK, since 2006. He is currently Deputy Director (Access) with responsibility for its online catalogue and promoting digital resource use among various user communities. He manages a number of major international research projects for the ADS, including projects on the archaeological use of Linked Open Data, data access and interoperability between archives and other scholarly resources. He has published extensively on digital preservation and reuse in archaeology, including new ways of sharing archaeological data and the impact this has on archaeological practice. Previous roles include senior posts with the West of Scotland Archaeology Service as well as with projects on the development of data standards and data aggregation with national and international scope.